STRUCTURED SEMANTIC EVIDENCE FOR MANAGING UNCERTAINTY UNDER SPARSE ASL SENSING

Anonymous authors
Paper under double-blind review

ABSTRACT

Perception-driven communication systems often operate under uncertain evidence because sensing can be incomplete and recognition models can make errors. A single predicted sequence or confidence score therefore cannot reveal which parts of the intended meaning are actually supported. We frame this as a context-conditioned semantic evidence problem: given Top-K recognition hypotheses, their relative probabilities, and dialogue context, infer which semantic information is required, determined, uncertain, or missing in the current conversational turn. We introduce a slot-specific cross-attention model that jointly reasons over recognition hypotheses and context to construct a structured semantic evidence representation: a collection of per-slot records indicating which meaning dimensions are required and whether their current evidence is Determined, Uncertain, Missing, or NotRequired. Experiments on ASL as a challenging testbed show that both recognition evidence and dialogue context make complementary contributions to structured semantic evidence estimation, while poorer recognition quality is consistently associated with weaker semantic recovery. Controlled context interventions further improve English meaning recovery when the correct conversational context is available. These results support structured semantic evidence as an intermediate representation between uncertain recognition and downstream conversational reasoning.

1 INTRODUCTION

Perception-driven communication systems (e.g., American Sign Language (ASL)) increasingly rely on outcomes recovered from incomplete sensing and imperfect recognition. The core challenge is not only that recognition can be wrong, but that a downstream agent often cannot tell *which parts of the underlying meaning are actually supported by the available evidence*. A single predicted sequence or scalar confidence collapses this structure: it may indicate what was most likely observed, but not what is known, ambiguous, or missing (Kuhn et al., 2023; Farquhar et al., 2024). This raises a more fundamental question: *how should conversational meaning be represented when perception itself is uncertain?*

We study this problem through American Sign Language (ASL), where egocentric based wearable sensing provides a concrete instance of uncertain embodied perception. Egocentric wearable ASL systems can support everyday communication, but their sensing is inherently incomplete and recognition errors can alter meaning-critical content (Wu et al., 2016; Gu et al., 2022). For example, the observed gloss sequence Want Cancel Appointment Change Friday may omit the negation marker Not. Both "I want to cancel my appointment" and "I do *not* want to cancel" are linguistically plausible interpretations, but they differ in a meaning-critical component—negation; manual gloss alone can omit grammatical information that is otherwise carried by non-manual markers such as facial expression (Michael et al., 2011; Zhang et al., 2025). More generally, recognition hypotheses may disagree strongly at the token level while preserving the same communicative meaning, or appear confident while omitting a critical component such as negation, person, time, or action. Thus, recognition uncertainty alone does not reveal what semantic information remains reliable.

We frame this as a *context-conditioned semantic evidence* problem. We use a *semantic slot* to denote a meaning dimension such as person, time, negation, action, object, location, or number. For each slot, a *per-slot evidence state* indicates whether that meaning dimension is Determined,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Uncertain, Missing, or NotRequired given the current recognition evidence and dialogue context. The collection of these slot-specific records forms the *structured semantic evidence representation*. Given Top-K recognition hypotheses, their relative probabilities, and the current dialogue context, the goal is to infer this representation for the current conversational turn. Recognition hypotheses describe what may have been perceived; dialogue context specifies what information matters in the current turn (Liao et al., 2021). We instantiate this representation with a learned slot-specific cross-attention model, in the spirit of sparse local slot attention for dialogue state tracking (Yang et al., 2023), that jointly attends to Top-K gloss hypotheses, recognition probabilities, and conversational context (§3.4). The resulting state explicitly represents *what is known, what remains unresolved, and what is not required*, rather than collapsing uncertainty into a single confidence score.

To evaluate this representation, we construct a signer-disjoint, human-labeled benchmark from real recognition outputs. Each example contains Top-K recognition hypotheses, conversational context, a human judgment of whether the available meaning is safe to communicate, and slot-level annotations describing required semantic information and where evidence remains insufficient. This benchmark supports two complementary questions. At the global level, release-risk prediction asks whether the current evidence is sufficient to communicate the utterance safely (Geifman & El-Yaniv, 2017; Xin et al., 2021; Srinivasan et al., 2024). At the semantic level, structured evidence estimation asks the finer question of what information is actually supported or unresolved. The benchmark also provides repair targets and acceptable actions for downstream action selection.

Our experiments support this formulation. On unseen signers, removing either recognition hypotheses or dialogue context substantially degrades structured semantic evidence estimation, showing that the learned representation depends jointly on recognition and contextual evidence. Across constructed recognition-quality Test subsets, the lower-WER cohort exhibits stronger per-slot evidence-state estimation than the higher-WER cohort, supporting an association between recognition quality and downstream semantic evidence. Controlled context interventions further show that correct dialogue context improves downstream meaning recovery relative to absent or mismatched context. Global release-risk detection provides a complementary signal for whether an utterance may be unsafe, while the structured state addresses the more informative question of which semantic content remains unresolved. Offline repair experiments further show that identifying what is uncertain and deciding how to repair it are distinct problems.

Contributions. (1) We formulate uncertain perception in communication as a context-conditioned semantic-evidence problem, where an agent must represent not only candidate meaning but also which semantic information is required, supported, uncertain, or missing.

(2) We introduce a slot-specific cross-attention model that constructs structured semantic evidence from Top-K recognition hypotheses, their probabilities, and dialogue context.

(3) We introduce a signer-disjoint, human-labeled communication-safety benchmark built from real recognition outputs, supporting evaluation of both global release risk and localized semantic uncertainty, together with repair targets for downstream analysis.

(4) Through five-seed signer-disjoint evaluation, recognition-quality analysis, input ablations, and controlled context interventions, we show that semantic evidence depends jointly on recognition hypotheses and conversational context, and that structured semantic evidence estimation and action selection are separable problems.

2 Problem Formulation

Input hypotheses and dialogue context. At conversational turn t, the system receives the partner query and dialogue history C_t. When signer input is required (§3.2), a signer video is decoded into a Top-K set of gloss hypotheses $\mathcal{H}_t = \{(g_k, p_k)\}_{k=1}^{K}$, where g_k is the kth gloss-sequence hypothesis and $p_k = \mathrm{softmax}(-\mathrm{beam_score}_k)$ denotes relative support within the retained Top-K list, rather than calibrated global correctness. The dialogue context specifies what semantic information the current turn requires, while the recognition hypotheses describe what may have been perceived.

Semantic slots and required information. We represent conversational meaning using seven semantic slot types: person, time, negation, action, object, location, and number. For each turn, the

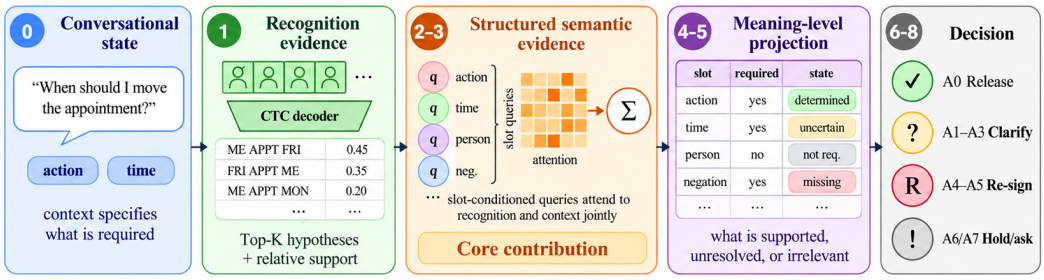

Figure 1: Overall pipeline from dialogue and noisy recognition to structured semantic evidence. The central representation assigns each semantic slot one of four evidence states: Determined, Uncertain, Missing, or NotRequired, before downstream meaning projection and action selection.

dialogue context determines which slots are required for the response. Recognition hypotheses provide candidate values and evidence for these same slots.

Context-conditioned semantic evidence. Our goal is to infer a structured semantic evidence representation

$$\mathcal{E}_t = f_\theta(\mathcal{H}_t, C_t) = \{\mathcal{E}_{t,s}\}_{s=1}^{S}, \tag{1}$$

where each semantic slot s has a slot-specific evidence record

$$\mathcal{E}_{t,s} = \big(r_{t,s}, P_{t,s}(v), P_{t,s}(z), p_{t,s}^{\text{repair}}, H_{t,s}\big). \tag{2}$$

Here, $r_{t,s}$ indicates whether slot s is required in the current turn, $P_{t,s}(v)$ represents support over candidate semantic values, and $P_{t,s}(z)$ is a distribution over the four per-slot evidence states

$$z_{t,s} \in \{\textsc{Determined}, \textsc{Uncertain}, \textsc{Missing}, \textsc{NotRequired}\}. \tag{3}$$

The term $p_{t,s}^{\text{repair}}$ denotes the predicted probability that slot s requires repair, and $H_{t,s}$ contains its supporting recognition hypotheses and attended dialogue-context evidence.

Thus, a semantic slot specifies *which meaning dimension is being evaluated*, a per-slot evidence state describes *how well that dimension is currently supported*, and $\mathcal{E}_t$ collects all slot-specific records into the structured semantic evidence representation consumed by downstream release-risk estimation and repair selection.

3 Method

3.1 System Overview

Figure 1 summarizes the framework. Stage 0 uses the partner turn and dialogue history to decide whether a signer response is needed. If so, Stage 1 converts sparse sign video into a Top-K set of gloss hypotheses with within-list relative support. Stages 2–3 form the core module: learned slot queries jointly attend to the recognition hypotheses, their probabilities, and dialogue context, producing a structured semantic evidence representation that assigns each slot one of four evidence states: Determined, Uncertain, Missing, or NotRequired. Stages 4–5 project the hypotheses into English candidate meanings and merge meaning-equivalent candidates. Stages 6–8 use the resulting evidence to release the meaning or select a source-appropriate intervention. The architecture preserves uncertainty until after semantic aggregation instead of committing to the 1-best gloss before conversational reasoning.

3.2 Conversation-level management (Stage 0).

A two-layer DistilRoBERTa router uses the partner query and dialogue context to decide whether a signer response is required. The first layer separates Signer-Response-Required from System-Handled; the latter is further classified as `direct-answer`, `ask`, or `refuse`, following ISO 24617-2 (Malchanau et al., 2024). The `act` route invokes the signer and Stage 1, whereas the

remaining routes answer, clarify, or request re-entry without sign recognition. Speaker identifiers are removed, contexts are grouped before partitioning, and model selection uses held-out validation data. We prioritize signer-required recall and evaluate domain shift on a separately authored 63-domain Test set (§4.1). Additional routing definitions and data-partition details are provided in Appendix A.The dataset is partitioned into three signer-disjoint folds. The detailed sample distribution across training, validation (Dev), and test sets for each fold is summarized in Table 5.

3.3 Perceptual Evidence Acquisition (Stage 1)

For translation, the signing video is first encoded by a ResNet-18 backbone, Temporal Lift Pooling, and a two-layer bidirectional LSTM with CTC beam search; full architecture details are in Appendix C.6. The training table contains 3,430 quality-controlled utterances derived from raw ASL transcripts, with matching, signer correction, and gloss canonicalization documented in Appendix C.7. For WER, the decoder returns its 1-best hypothesis. For Stage 2, we retain up to five unique CTC beams and compute their within-list support as $p_k = \mathrm{softmax}(-\mathrm{score}_k)$. Beams that collapse to the same gloss sequence are merged by summing their normalized mass. The resulting Top-K hypotheses, token confidences, and timesteps provide recognition evidence for downstream semantic analysis. Low-recall meaning-critical glosses, particularly for person, negation, action, and object, motivate the slot reliability prior in §3.4.

3.4 Cross-Attentive Semantic Evidence (Stage 2 & 3)

Stage 2: initial semantic parsing and context-slot extraction. The Top-K gloss candidates are parsed into the seven semantic slots (§2) using a hand-authored, interpretable lexicon (gloss $\rightarrow$ slot). Independently, and without access to the signer's ground-truth answer, the partner query and dialogue history are parsed into the required context slots that the current turn's response must address. Each slot's reliability is estimated once from non-Test recognition statistics and frozen before final evaluation:

$$\text{reliability}(s) = \min_{g \in \text{anchors}(s)} \text{Recall}(g), \qquad \text{risk_prior}(s) = 1 - \text{reliability}(s), \tag{4}$$

using $\min$ rather than mean, since a single low-recall anchor gloss should dominate the slot's reliability regardless of other anchors. This static prior is supplied to Stage 3 alongside the per-utterance evidence below.

Stage 3: slot-conditioned evidence fusion. Figure 2 details the core computation. For each slot s, a learned query q_s attends separately to Top-K gloss tokens and dialogue-context tokens. Candidate identity, rank, and relative beam probability are embedded with each recognition token. A learned modality gate fuses the slot query, recognition evidence, and contextual evidence. Three multitask heads output the probability that the slot is required, a distribution over four per-slot evidence states, and the probability that the slot requires repair. The resulting structured semantic evidence representation also retains the candidate values V_s extracted by the semantic parser, their supporting hypotheses, attended context cues, and the fixed slot reliability prior. For each semantic slot s, the model selects the predicted evidence state $z_{t,s}$ from the distribution $P_{t,s}(z)$ defined in §2. For example, the model may predict person $= \{\text{I} : 0.96\}$ as DETERMINED, while negation $= \{\text{Positive} : 0.53, \text{Negative} : 0.47\}$ is UNCERTAIN. Required slots not in state DETERMINED are retained for repair at Stage 8. The selected model represents recognition uncertainty through Top-K hypotheses and probabilities; raw TemporalConv attention is evaluated as an input ablation. Unlike a sentence-level confidence score, this representation preserves the location and type of uncertainty. It can therefore distinguish an utterance with one unresolved required slot from one with several coupled deficiencies, even when their aggregate risk is similar. The representation serves as a common interface between recognition, meaning projection, and downstream repair.

3.5 Meaning Projection and Meaning-Level Uncertainty(Stages 4–5)

Stage 4: context-conditioned English meaning projection. Each Top-K gloss candidate is first converted by the fixed semantic parser into a candidate semantic frame. The candidate glosses, their relative probabilities, semantic frames, and partner context are then projected into English sentence candidates using a fixed GPT-5.5 prompt. For each candidate, the projection returns an English

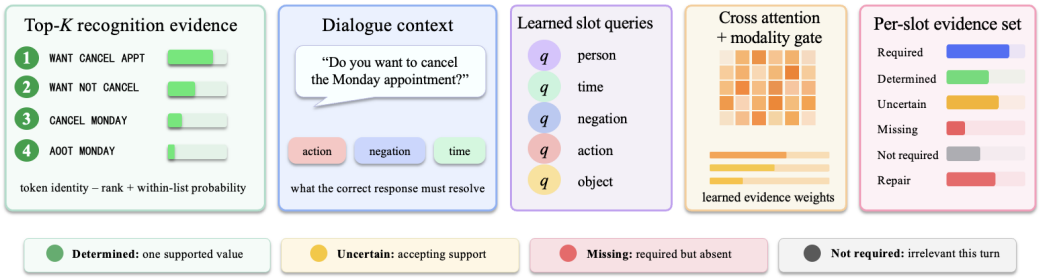

Figure 2: Core slot-specific cross-attention module. Each semantic slot provides a learned query; Top-K recognition tokens carry candidate identity, rank, and within-list probability, while context provides turn-specific information requirements. Cross-attention and learned gating produce required-slot probability, four-way per-slot evidence-state probabilities, and repair probability for every semantic slot.

sentence, a structured meaning key, a projection-coverage score, and any semantic components that could not be expressed. For example, the gloss candidates ME LIKE APPLE and LIKE APPLE ME may both project to "I like apples." Reference English, safety labels, and action labels are withheld during generation and joined only afterward for evaluation.

The learned Stage 3 representation is evaluated separately from the English projector: it supplies localized required-slot, evidence-state, and repair predictions for evidence-sufficiency analysis and repair targeting, whereas Stage 4 evaluates whether the candidate recognition hypotheses can be projected into contextually appropriate English meanings.

Stage 5: meaning-level uncertainty. English sentence candidates are clustered by meaning equivalence. Writing p_k for the Top-K relative confidence of the gloss candidate underlying English candidate k, each meaning cluster m receives posterior mass $P(m) = \sum_{k \in m} p_k$, and cluster-level risk is

$$R_{\text{cluster}} = 1 - \max_m P(m), \tag{5}$$

i.e., the probability mass *not* concentrated on the single most likely meaning – low when different Top-K candidates converge to the same meaning despite gloss-level disagreement, high when they diverge. This stage determines whether different gloss hypotheses still converge to the same communicative meaning, independent of whether individual glosses agree.

3.6 EVIDENCE SUFFICIENCY AND DOWNSTREAM INTERACTION PROTOTYPE: STAGES 6–8

Stage 6: evidence sufficiency. We evaluate evidence sufficiency at two complementary levels. For utterance-level unsafe-meaning detection, we fit a class-balanced logistic combiner on development data using runtime-available recognition and semantic-risk summaries. Its input vector is

$$\mathbf{x}_t = [R_{\text{mean}}, R_{\text{min}}, R_{\text{top1}}, H_{\text{beam}}, R_{K=3}, R_{K=5}, R_{\text{existing}}, R_{\text{combined}}], \tag{6}$$

where the components summarize sentence confidence, Top-1 support, beam entropy, and semantic disagreement. The resulting global risk is

$$R_{\text{global}} = \sigma(\mathbf{w}^\top \mathbf{x}_t + b), \qquad P(\text{sufficient}) = 1 - R_{\text{global}}. \tag{7}$$

This score estimates whether the current utterance is safe to release. In parallel, the Stage 3 multi-task heads provide localized required-slot, evidence-state, and repair predictions that identify which semantic information remains unresolved. Thus, global risk supports the release-versus-intervene decision, while structured per-slot evidence supports deficiency attribution and repair targeting.

Figure 1 shows these complementary downstream uses. Evidence judged sufficient follows $A0$. Otherwise, the interaction prototype uses the localized evidence state to attribute the deficiency as semantic, visual, or contextual and select a corresponding intervention.

Return paths. Semantic repairs return to Stage 6, visual repairs return to Stage 1, and partner clarification returns to Stage 2. Unlike Stage 0 `ask`, which occurs before signer involvement, $A7$ is triggered after semantic analysis exposes insufficient conversational context.

4 Experiments and Results

4.1 Experimental Setup

Stage 0 and Stages 1–2 use separate datasets because they address different pipeline decisions. Stage 0 uses 3,071 partner queries, including 826 signer-required examples, partitioned by grouped dialogue context. Conditional on the signer-response route, Stage 1 is evaluated on a signer-partitioned ASL corpus containing 3,430 quality-controlled utterances from 19 signers. Therefore, the offline Stage 2 benchmark represents utterances that have already entered the signer-response branch; it is not a one-to-one continuation of the Stage 0 examples.

For semantic-evidence evaluation, cross-fold recognition outputs are aligned by `VideoName` with dialogue context, reference English and gloss, and human semantic annotations. After excluding incomplete matches, the resulting cohort contains 3,335 utterances: 1,920 for development and 1,415 for final testing from 10 unseen signers. Each Test utterance contains up to five gloss hypotheses with relative probabilities and labels for seven semantic slots, yielding 9,905 slot decisions. The Test set contains 1,339 UNSAFE and 76 SAFE utterances. Model selection and threshold fitting use only non-Test data, and per-slot results are reported over five fixed seeds.

We evaluate whether (i) Top-K recognition evidence and dialogue context jointly improve structured semantic evidence estimation (Table 1), (ii) recognition quality is associated with semantic recovery on constructed lower- and higher-WER Test subsets, and (iii) correct context improves English meaning projection relative to absent or shuffled context (Table 3). We additionally report global unsafe-meaning detection and offline action selection.

4.2 Routing Performance (Stage 0).

As an upstream routing component, the fine-tuned DistilRoBERTa model achieves 99.78% signer/system routing accuracy on the original 457-example held-out split. On a separately constructed 500-example cross-domain Test set, the domain-generalized variant achieves 80.40% accuracy, 87.07% balanced accuracy, 100% signer-required recall, and 74.14% system-handled recall. Most errors occur in previously unseen direct-answer scenarios, while signer-required recall remains at 100%.

4.3 Structured Semantic Evidence Estimation

Overall performance. Table 1 directly tests the inputs in the proposed representation. The complete recognition-plus-context model reaches 0.956 required-slot AUROC, 0.660 evidence-state macro-F1, and 0.927 repair AUROC. Its evidence-state F1 is 0.255 above the slot-prior control. Removing recognition hypotheses reduces F1 by 0.139, while removing context reduces it by 0.158. A zero-shot GPT-5.5 baseline given the same Top-K hypotheses, probabilities, context, and slot schema reaches 0.503 state macro-F1, 0.157 below the proposed model. These results indicate that both recognition evidence and dialogue context are important, and that directly prompting a general-purpose LLM does not recover the structured per-slot evidence performance of the trained estimator. The two input sources are complementary rather than interchangeable: recognition hypotheses provide candidate semantic content, whereas context identifies which content is required in the current turn. Their joint gain supports the proposed context-conditioned evidence formulation rather than either an ASR-confidence-only or context-only alternative.

Top-K evidence and raw visual features. Holding the semantic projection fixed, disagreement-based unsafe-meaning AUROC is 0.500 for $K{=}1$, 0.713 for $K{=}3$, and 0.719 for $K{=}5$; AURC decreases from 0.955 to 0.912 and 0.911. The single-hypothesis setting cannot expose between-candidate semantic disagreement, while most of the gain is recovered by three hypotheses. Adding raw TemporalConv features to the per-slot model yields 0.600 evidence-state F1 and 0.906 repair AUROC, below the selected model's 0.660 and 0.927. Under the available slot supervision, Stage 1

Table 1: Structured semantic evidence estimation on ten unseen Test signers (mean±std over five seeds). State Acc. and State F1 evaluate four-way per-slot evidence-state prediction; ΔF1 is measured relative to the complete Top-K recognition-plus-context model. GPT-5.5 direct is a single zero-shot run using the same Top-K hypotheses, probabilities, context, and slot schema, without reference annotations.

Input	Req. AUROC	State Acc.	State F1	ΔF1	Rep. AUROC
Slot prior only	.871±.002	.565±.042	.405±.034	−.255	.858±.002
Context only	.930±.003	.637±.017	.521±.012	−.139	.898±.008
Top-K recognition only	.895±.003	.586±.022	.502±.019	−.158	.866±.002
GPT-5.5 direct (zero-shot)	.865	.611	.503	−.157	.808
Top-K recognition + context	**.956±.007**	**.748±.024**	**.660±.028**	–	**.927±.005**

Table 2: Recognition-quality sensitivity on two constructed, signer-matched Test cohorts of 300 utterances each. Per-slot results are averaged over five frozen-model seeds. Absolute and relative differences compare Lower-WER against Higher-WER, using Higher-WER as the reference for relative differences.

Metric	Lower	Higher	Abs. diff.	Rel. diff.
Utterances	300	300	–	–
Micro-WER (%) ↓	47.45	87.84	−40.39 pp	46.0%
Required-slot AUROC ↑	.979	.970	+.009	+1.0%
Required-slot accuracy ↑	.940	.925	+.015	+1.6%
Evidence-state accuracy ↑	.782	.771	+.011	+1.4%
Evidence-state macro-F1 ↑	.685	.681	+.004	+0.5%

hypotheses and probabilities carry more transferable evidence than an additional unaligned temporal feature stream.

Recognition-quality sensitivity. To validate the structured semantic evidence under different recognition quality, we constructed and compared two 300-utterance sub-test cohorts without utterance overlap, as shown in Table 2. Their micro-WER differs substantially (47.45% versus 87.84%). Across five frozen-model seeds, the lower-WER cohort improves required-slot AUROC from 0.970 ± 0.007 to 0.979 ± 0.003, required-slot accuracy from 0.925 ± 0.010 to 0.940 ± 0.009, and evidence-state accuracy from 0.771 ± 0.026 to 0.782 ± 0.019. Evidence-state macro-F1 changes more modestly, from 0.681 ± 0.031 to 0.685 ± 0.023. Because the cohorts are selected using observed sentence WER, this stress test establishes an association between recognition quality and structured semantic recovery rather than a natural-distribution or causal estimate.

4.4 Controlled Context Intervention

We run a pinned GPT-5.5 meaning-projection prompt on all 1,415 Test utterances with the correct partner context, no context, or a context shuffled from another utterance. Generation receives the Top-K hypotheses and their relative probabilities, while reference English is used only for evaluation. Correct context improves Token-F1 by 0.038 over no context (paired-bootstrap 95% CI $[0.032, 0.044]$) and by 0.039 over shuffled context ($[0.033, 0.045]$). The corresponding chrF gains are 0.026 ($[0.022, 0.031]$) and 0.027 ($[0.023, 0.031]$). The near-identical no-context and shuffled-context results localize the improvement to relevant conversational evidence rather than additional prompt text. Absolute reconstruction remains difficult under the noisy Stage 1 outputs, while the paired intervention isolates a consistent context contribution.

4.5 Complementary Safety and Action Analysis

Global unsafe-meaning detection. This utterance-level task complements the localized per-slot evidence-state task. Sentence-mean recognition risk reaches 0.771 AUROC and 0.895 AURC; beam entropy reaches 0.732/0.897, and Top-5 semantic disagreement reaches 0.719/0.911. A class-balanced available-evidence fusion fitted on validation data improves these results to 0.792 AUROC

Table 3: Controlled GPT-5.5 context intervention on 1,415 Test utterances. The English reference is held out during generation.

Context	Exact (%) ↑	Token-F1 ↑	chrF ↑	BLEU-4 ↑
Correct	**1.41**	**.278**	**.256**	**6.12**
None	.57	.240	.230	3.42
Shuffled	.49	.239	.229	3.38
Correct − None	+.85	+.038	+.026	+2.70
Correct − Shuffled	+.92	+.039	+.027	+2.74

Table 4: Complementary unsafe-meaning detection on the complete human-labeled Test set.

Risk signal	AUROC ↑	AURC ↓	ECE ↓
Sentence-mean recognition risk	.771	.895	.474
Beam entropy	.732	.897	**.054**
Top-5 semantic disagreement	.719	.911	.881
Learned available-evidence fusion	**.792**	**.886**	.294

and 0.886 AURC. Its ECE is 0.294, compared with 0.474 for sentence-mean risk and 0.054 for beam entropy. Multi-source fusion therefore improves ranking, while beam entropy remains the best-calibrated individual signal.

Offline action selection. The current $A0$–$A7$ heuristic recalls 96.71% of unsafe utterances for repair, with 3.89% release coverage and 55.59% release/repair balanced accuracy. Exact agreement with the human action is 7.35%, increasing to 11.17% when human-marked acceptable alternatives are included. These results separate evidence estimation from action optimization: the system localizes unresolved evidence substantially more accurately than it selects the preferred intervention. The benchmark contains a single turn per utterance, so the action analysis ends at offline agreement; post-action semantic recovery requires the response elicited by $A1$–$A3/A7$ or a new sign video elicited by $A4/A5$.

5 Related Work and Discussion

Semantic uncertainty and structured evidence representation. A single sequence-level or token-level confidence score does not reveal which part of a meaning is actually supported; semantic uncertainty instead aggregates probability mass over meaning-equivalence classes (Kuhn et al., 2023; Farquhar et al., 2024). Subsequent work refines what that aggregation should operate over: kernel methods represent pairwise semantic similarity among candidates directly, estimating uncertainty via a von Neumann entropy over the resulting similarity structure (Nikitin et al., 2024); confidence can instead be read off local density in a semantic embedding space (Qiu & Miikkulainen, 2024); semantic uncertainty can be encoded in a model's own latent representations and recovered with a lightweight probe from a single forward pass, rather than requiring multiple sampled generations (Han et al., 2024; Kossen et al., 2024); and pairwise semantic similarity between candidates further sharpens cluster-based semantic entropy (Nguyen et al., 2025). A related line argues that the candidate set itself should be constructed to cover genuinely distinct semantic alternatives, rather than accepting whichever near-duplicate hypotheses a decoder happens to produce (Park & Cho, 2026) – a concern that applies directly to our Top-K beam candidates. Li et al. (2025) argue more broadly that distinct decision questions (e.g., is this input unfamiliar? is this specific claim supported?) are routinely collapsed into a single score, which motivates our separation of global release-risk detection from localized per-slot evidence estimation (§3.6–§3.6). Probabilistic, rather than deterministic, representations of ambiguous multimodal input have also been explored for composed image retrieval (Tang et al., 2025) and, closest to our modality, for uncertainty-aware sign-language video retrieval (Wu et al., 2024). All of this work characterizes uncertainty over a model's own generated outputs or embeddings; none of it asks, for an embodied communication channel, *which required piece of meaning remains unsupported*, or conditions that judgment jointly on upstream perceptual evidence and the current dialogue turn. We adopt slot-conditioned atten-

tion from dialogue-state tracking, which tracks belief state incrementally across turns (Liao et al., 2021) using sparse, slot-specific attention (Yang et al., 2023), but condition each slot's evidence state jointly on Top-K recognition hypotheses and dialogue context rather than on text-only dialogue history.

Calibration, selective prediction, and ambiguity classification. Structured calibration methods produce entity-level confidence without extra training cost (Jagannatha & Yu, 2020); selective prediction studies when a model should decline to answer (Xin et al., 2021; Gu & Hopkins, 2023), and ReCoVERR reduces over-abstention by first searching for corroborating evidence (Srinivasan et al., 2024). CLAMBER offers a taxonomy for classifying ambiguous information needs (Zhang et al., 2024), and ISO 24617-2 standardizes dialogue-act annotation (Malchanau et al., 2024). Most directly, Overman & Bayati (2025) calibrate a post-hoc arbitration threshold between a utility-maximizing Primary policy and a risk-bounding Guardian policy via conformal risk control, without requiring model logits. We calibrate at the slot level, matching Jagannatha & Yu's granularity, but substitute Stage 1's offline error analysis for a trained calibration model; ReCoVERR's evidence-gathering principle informs $A1$, CLAMBER informs our slot taxonomy, Stage 0's four labels are grounded in ISO 24617-2, and Overman & Bayati's Primary/Guardian arbitration is instantiated directly as Stage 7's $A0$-vs-intervention routing (§3.6). These methods assume the confidence signal originates within the deciding model; in ours, part of it (gloss recall) originates upstream in Stage 1 and must be propagated across a system boundary – a cross-system calibration problem this literature does not consider.

6 Limitations

The per-slot model is weakly supervised by one annotator; without frame-level labels, its attention weights are not validated temporal grounding. Raw TemporalConv features reduce rather than improve signer-disjoint performance, so the selected model uses Top-K hypotheses and relative probabilities as its recognition evidence. The Test set is also highly safety-skewed (94.63% unsafe), making macro metrics and AUROC essential and limiting prevalence-dependent deployment claims. English reconstruction from the aligned Test hypotheses remains difficult in absolute terms, despite the significant gain from correct dialogue context. Stage 0 remains conservative under domain shift: after augmentation it retains 100% signer-required recall but only 74.14% system-handled recall. Finally, $A0$–$A7$ is evaluated only through offline agreement and proxy cost. The current corpus contains no post-action responses or re-sign videos from which to measure recovery, latency, signer burden, or post-repair safety. The frame-slot lexicon is scenario-scoped, and the split measures unseen-signer rather than unseen-sentence generalization.

7 Conclusion

We reframe sparse-sensing ASL communication as uncertainty management rather than unconditional translation. On unseen signers, learned slot-specific cross-attention over Top-K recognition hypotheses, their relative probabilities, and dialogue context reaches 0.956 required-slot and 0.927 repair AUROC. Removing recognition evidence or context reduces evidence-state macro-F1 by 0.139 and 0.158, respectively. The controlled context intervention improves English Token-F1 by 0.038 over absent context and 0.039 over shuffled context. Together, these results establish structured semantic evidence as an interpretable interface between uncertain recognition and subsequent conversational decisions. The offline action results further identify action optimization as the next layer beyond structured semantic evidence estimation.

AI use statement

We used generative AI tools to assist with drafting and revising manuscript text. The authors take full responsibility for the accuracy of the resulting content: all reported quantitative results were produced and verified by the authors' own code and pipelines, and all citations and factual claims were checked against primary sources by the authors. This disclosure follows the ICLR 2027 policies on LLM usage.

Ethics statement

This work does not involve new data collection by the authors and therefore did not require its own IRB review. All experiments reuse an existing egocentric ASL corpus collected under a separate approved research protocol (Appendix C), which was gathered under that work's own approved Institutional Review Board (IRB) protocol, with informed consent from all Deaf participants. Capture used an egocentric viewpoint that excludes the signer's face, mitigating some privacy risk relative to third-person recordings, though hand appearance and signing content remain potentially identifying. Our partner-clarification mechanism ($A7$, §3.6) is scoped to context-alignment failures rather than open-ended delegation, limiting the range of signer content a communication partner is asked to weigh in on. We have not conducted human-subject evaluation of the repair policy itself; any future human validation with ASL-knowledgeable raters would require its own IRB review.

Reproducibility statement

The Method section (§3) and Appendix C document the recognition architecture, training data construction pipeline (including the canonicalization quality-control procedure, Table 6), and the signer-disjoint split methodology in sufficient detail to reproduce the Stage 1/Stage 0 pipeline. We do not release an anonymized code repository alongside this submission; the architectural, training, and data-construction details in §3 and Appendix C are intended to be sufficient for reproduction from these descriptions. The underlying ASL corpus (Appendix C) is not yet public; a complete description of its collection and annotation protocol is provided there for completeness and self-containedness of this submission. An automated audit confirmed zero target-answer leakage into partner dialogue context across all 3,430 processed rows of the underlying map (30/30 integrity tests passed), supporting the validity of the semantic-level evaluation methodology in §4.1.

References

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630, 2024. URL https://api.semanticscholar.org/CorpusID:270615909.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks, 2017. URL https://arxiv.org/abs/1705.08500.

Yutong Gu, Chao Zheng, Masahiro Todoh, and Fusheng Zha. American sign language translation using wearable inertial and electromyography sensors for tracking hand movements and facial expressions. *Frontiers in Neuroscience*, 16, 2022. URL https://api.semanticscholar.org/CorpusID:250646227.

Zhengyao Gu and Mark Hopkins. On the evaluation of neural selective prediction methods for natural language processing. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7888–7899, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.437. URL https://aclanthology.org/2023.acl-long.437/.

Jiatong Han, Jannik Kossen, Muhammed Razzak, and Yarin Gal. Semantic entropy neurons: Encoding semantic uncertainty in the latent space of LLMs. In *MINT: Foundation Model Interventions Workshop*, 2024. URL https://openreview.net/forum?id=BbZKxrZCNn.

Abhyuday Jagannatha and Hong Yu. Calibrating structured output predictors for natural language processing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2078–2092, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.188. URL https://aclanthology.org/2020.acl-main.188/.

Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth Malik, and Yarin Gal. Semantic entropy probes: Robust and cheap hallucination detection in LLMs, 2024. URL https://arxiv.org/abs/2406.15927.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation, 2023. URL https://arxiv.org/abs/2302.09664.

Yucen Lily Li, Daohan Lu, Polina Kirichenko, Shikai Qiu, Tim G. J. Rudner, C. Bayan Bruss, and Andrew Gordon Wilson. Position: Supervised classifiers answer the wrong questions for OOD detection. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025. URL https://openreview.net/forum?id=UXZJ3aL8vE.

Lizi Liao, Le Hong Long, Yunshan Ma, Wenqiang Lei, and Tat-Seng Chua. Dialogue state tracking with incremental reasoning. *Transactions of the Association for Computational Linguistics*, 9: 557–569, 2021. doi: 10.1162/tacl_a_00384. URL https://aclanthology.org/2021.tacl-1.34/.

Andrei Malchanau, Volha Petukhova, and Harry Bunt. Fusing ISO 24617-2 dialogue acts and application-specific semantic content annotations. In *Proceedings of the Joint Workshop on Interoperable Semantic Annotation (ISA-20)*, pp. 122–132, May 2024. URL https://aclanthology.org/2024.isa-1.15/.

Nicholas Michael, Peng Yang, Qingshan Liu, Dimitris Metaxas, and Carol Neidle. A framework for the recognition of nonmanual markers in segmented sequences of american sign language. In Jesse Hoey, Stephen McKenna, and Emanuele Trucco (eds.), *Proceedings of the British Machine Vision Conference*, pp. 124.1–124.12. BMVA Press, 2011. ISBN 1-901725-43-X. http://dx.doi.org/10.5244/C.25.124.

Dang Nguyen, Ali Payani, and Baharan Mirzasoleiman. Beyond semantic entropy: Boosting LLM uncertainty quantification with pairwise semantic similarity, 2025. URL https://arxiv.org/abs/2506.00245.

Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. Kernel language entropy: Fine-grained uncertainty quantification for LLMs from semantic similarities, 2024. URL https://arxiv.org/abs/2405.20003.

William Overman and Mohsen Bayati. Conformal arbitrage: Risk-controlled balancing of competing objectives in language models. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 70305–70337. Curran Associates, Inc., 2025. doi: 10.52202/085713-2364. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/65a655c5a267f678fd3e897e4137ef53-Paper-Conference.pdf.

Ji Won Park and Kyunghyun Cho. Efficient semantic uncertainty quantification in language models via diversity-steered sampling, 2026. URL https://arxiv.org/abs/2510.21310.

Xin Qiu and Risto Miikkulainen. Semantic density: Uncertainty quantification for large language models through confidence measurement in semantic space, 2024. URL https://arxiv.org/abs/2405.13845.

Tejas Srinivasan, Jack Hessel, Tanmay Gupta, Bill Yuchen Lin, Yejin Choi, Jesse Thomason, and Khyathi Chandu. Selective "selective prediction": Reducing unnecessary abstention in vision-language reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 12935–12948, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.767. URL https://aclanthology.org/2024.findings-acl.767/.

Haomiao Tang, Jinpeng Wang, Yuang Peng, Guanghao Meng, Ruisheng Luo, Bin Chen, Long Chen, Yaowei Wang, and Shu-Tao Xia. Modeling uncertainty in composed image retrieval via probabilistic embeddings. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1210–1222, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.61. URL https://aclanthology.org/2025.acl-long.61/.

Jian Wu, Lu Sun, and Roozbeh Jafari. A wearable system for recognizing american sign language in real-time using imu and surface emg sensors. *IEEE Journal of Biomedical and Health Informatics*, 20(5):1281–1290, 2016. doi: 10.1109/JBHI.2016.2598302.

Xuan Wu, Hongxiang Li, Yuanjiang Luo, Xuxin Cheng, Xianwei Zhuang, Meng Cao, and Keren Fu. Uncertainty-aware sign language video retrieval with probability distribution modeling, 2024. URL https://arxiv.org/abs/2405.19689.

Ji Xin, Raphael Tang, Yaoliang Yu, and Jimmy Lin. The art of abstention: Selective prediction and error regularization for natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1040–1051, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.84. URL https://aclanthology.org/2021.acl-long.84/.

Longfei Yang, Jiyi Li, Sheng Li, and Takahiro Shinozaki. Dialogue state tracking with sparse local slot attention. In Yun-Nung Chen and Abhinav Rastogi (eds.), *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, pp. 39–46, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.nlp4convai-1.4. URL https://aclanthology.org/2023.nlp4convai-1.4/.

Han Zhang, Rotem Shalev-Arkushin, Vasileios Baltatzis, Connor Gillis, Gierad Laput, Raja Kushalnagar, Lorna Quandt, Leah Findlater, Abdelkareem Bedri, and Colin Lea. Towards ai-driven sign language generation with non-manual markers, 2025. URL https://arxiv.org/abs/2502.05661.

Tong Zhang, Peixin Qin, Yang Deng, Chen Huang, Wenqiang Lei, Junhong Liu, Dingnan Jin, Hongru Liang, and Tat-Seng Chua. CLAMBER: A benchmark of identifying and clarifying ambiguous information needs in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10746–10766, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.578. URL https://aclanthology.org/2024.acl-long.578/.

A Stage 0 Routing Details

Stage 0 uses a two-layer DistilRoBERTa router. The first layer predicts whether the current partner turn requires a response from the signer. If signer involvement is required, the label is `act`, and the utterance proceeds to Stage 1. Otherwise, the second layer selects one of three system-handled outcomes: `direct-answer` returns an existing system response, `ask` requests missing information from the partner, and `refuse` requests that the partner re-enter an invalid or unsupported query.

The distinction between Stage 0 `ask` and action $A7$ is temporal: `ask` occurs before signer involvement when the partner query is incomplete, whereas $A7$ occurs after sign recognition when the recognized semantic evidence cannot be aligned reliably with the conversational context.

Speaker identifiers are removed from the router input, and normalized dialogue contexts are grouped before partitioning to prevent the same context template from appearing in both model development and final evaluation. Model and threshold selection use held-out validation data. Domain generalization is evaluated on a separately authored 63-domain, 500-example Test set that is not used for fine-tuning.

Table 5: Dataset partition statistics across three signer-disjoint folds.

Fold	Train	Dev	Test
Fold 1	2,265	521	644
Fold 2	2,118	638	674
Fold 3	2,187	904	339

B EVPI-INSPIRED MINIMAL-REPAIR OBJECTIVE

When Stage 7 routes an utterance to Guardian, the unresolved required slots are

$$S_{\text{repair}} = \{s \in S_{\text{required}} \mid \text{state}(s) \neq Determined\}. \tag{8}$$

For each candidate intervention a, the policy estimates the slots the action can recover, its residual semantic risk $R_{\text{after}}(a)$, and its cost $C(a)$. The cost combines expected risk reduction, user interaction burden, and repair latency. Action selection is written as

$$a^* = \arg \min_{a \in \{A1,\ldots,A7\}} C(a) \quad \text{s.t.} \quad S_{\text{repair}} \subseteq \text{RecoverableSlots}(a), \;\; R_{\text{after}}(a) \leq \alpha. \tag{9}$$

This objective is EVPI-inspired: an action is useful only when the expected information it obtains reduces the unresolved semantic risk enough to justify its interaction cost. Evidence-deficiency attribution restricts the feasible actions before optimization: semantic conflict routes to $A1$–$A3$, visual insufficiency to $A4$–$A5$, contextual ambiguity to $A7$, and the absence of any feasible risk-reducing action to $A6$.

C UNDERLYING EGOCENTRIC ASL CORPUS

Our Stage 1 recognizer is trained on data drawn from an egocentric stereo ASL corpus collected as part of concurrent, separately documented work; the corpus itself is not yet public, and is described here, without external citation, for completeness and self-containedness of the present submission.

C.1 CAPTURE SETUP

Signing was recorded with a head-mounted stereo RGB camera capturing first-person hand motion, alongside multiple calibrated, synchronized RGB-D cameras placed around the signer to provide multi-view coverage for 3D ground-truth reconstruction. Capture frame rate was varied across sessions to increase representational diversity.

C.2 SIGNERS AND PROTOCOL

Deaf, native ASL-using participants took part under a watch-and-repeat protocol: each participant viewed a short reference ASL expression and then reproduced it from memory. Sessions ran approximately two hours with scheduled breaks, during which participants removed and re-mounted the headset – a deliberate choice to capture natural variability in device positioning rather than an idealized, fixed viewpoint.

C.3 LINGUISTIC ANNOTATION

Certified ASL specialists segmented the continuous recordings into sentence-level clips and annotated each with a citation-form ASL gloss and an aligned English translation, following a two-pass protocol (initial annotation by one specialist, verification by a second, more experienced specialist). Annotator uncertainty was recorded alongside each segment rather than discarded silently.

C.4 3D HAND ANNOTATION

A semi-automated pipeline produced frame-accurate 3D hand keypoints: an off-the-shelf landmark detector proposed initial 2D joints per camera view, human annotators corrected them, depth was assigned via local neighborhood aggregation from the RGB-D stream, points were lifted to a shared 3D coordinate frame via calibrated camera intrinsics/extrinsics, and multi-view estimates were fused with an explicit tolerance for cross-view disagreement, flagging ambiguous joints as unlabeled rather than guessing. A final frame-by-frame manual audit corrected residual anatomically implausible poses.

C.5 Relationship to Stage 1 Training Data

The Stage 1 training table (§3.3) is derived from a task-specific subset of this corpus's raw session-level transcripts, independently re-derived and quality-controlled rather than inherited from any pre-existing split of the corpus; the two should not be assumed to share the same signer count or split boundaries.

C.6 Stage 1 Sequence Model Architecture Detail

Each frame is encoded by the shared-weight ResNet-18 backbone into a 512-d feature, producing a $(B, 512, T)$ frame-wise feature sequence. A 1D temporal convolution module performs temporal modeling and downsampling using Temporal Lift Pooling: the sequence is split into even/odd-indexed sub-sequences, a learned predictor estimates the odd sub-sequence from the even one, and the residual is used to update the even sub-sequence, which serves as the downsampled low-frequency component; both the updated low-frequency component and the high-frequency residual are retained via a learned weighted combination, unlike standard max/average pooling. A bidirectional 2-layer LSTM then models long-range temporal context, followed by a weight-normalized linear classification head and a CTC beam-search decoder. We adopt a dual-CTC self-distillation training scheme: one CTC loss is computed on the temporal-convolution output directly, and a second on the BiLSTM output; the deeper branch's distribution (detached) is distilled into the shallow branch via KL divergence at temperature $T{=}8$.

C.7 Matching, Signer Correction, and Canonicalization Detail

Matching. 4,007 raw gloss-transcribed clips are matched against a pool of 1,007 candidate English sentences using TF-IDF cosine similarity with WordNet-synonym expansion, backed by a clustering-consensus fallback for low-confidence individual matches, combined with human-reviewed approval at scale.

Signer identity correction. Raw session identifiers (e.g., `Abdul_01_04`, `Abdul_02_25`) naively suggest 39 distinct signers, but in fact encode multiple recording sessions from the same 19 real individuals. Treating sessions as independent identities would leak a signer's signing style across data partitions; we extract true signer identity from the leading alphabetic run of each session string and split at the level of the resulting 19 real signers.

Motion-similarity outlier verification. Rather than manually flagging "unusual" signers by inspection, we use a hand-keypoint dynamic time warping (DTW) motion-similarity report to identify *true* global outliers: signers with both low average similarity and a low pass rate against every other signer, not merely a single counterpart. This distinguishes genuine outliers (forced into training) from signers that appeared anomalous only due to a one-off pairing mismatch, which are returned to the pool used for subsequent development and evaluation partitioning.

Canonicalization quality control. Each translation sentence must map to exactly one canonical gloss string for training-class consistency. We canonicalize by majority vote over cleaned gloss strings within each translation group, falling back to the longest/most-complete variant when no majority exists. This surfaced several classes of contamination, summarized in Table 6 (Appendix D), which we address with a standing rule – a row is retained only if its own transcript contains a substantive anchor word tying it to its assigned translation – and by always recomputing each canonical fresh from source transcripts after any group reassignment.

D Canonicalization Quality-Control Issues

Table 6: Issue categories surfaced during gloss-translation canonicalization (Appendix C.7), and the fix applied to each.

Issue	Symptom	Fix
Majority-vote contamination	Coincidentally identical sparse content wins a false majority, overwriting a group's true canonical	Every group member's gloss is refreshed from its own source transcript before recomputing the canonical; never trust a prior displayed value
Uninformative shortest-canonical	A tied vote selects an overly short variant, dropping key content words	Fallback changed from "pick shortest" to "pick longest/most complete"; canonicals ≤ 3 words relative to the richest group member are flagged for manual review
Unsupported "ride-along" members	A row is retained in a group despite its own content not supporting the assigned translation	A row is kept only if its own transcript contains a substantive anchor word tying it to the assigned translation; otherwise excluded
Regex coverage gaps	Structural variants of known markup patterns leak through as literal content	Regexes generalized; full table re-processed from source transcripts after each fix
Grammar-tag / referent mismatch	Content with incompatible question type or referent is merged into the same group	Every member is cross-checked against the candidate-sentence pool and reassigned per-row rather than kept under an ambiguous parent group